Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.046081 Published Online: 04 February 2024

ARTICLE

Restoration of the JPEG Maximum Lossy Compressed Face Images with Hourglass Block-GAN

Jongwook Si[1] and Sungyoung Kim[2,*]

[1]Department of Computer AI Convergence Engineering, Kumoh National Institute of Technology, Gumi, 39177, Korea

[2]Department of Computer Engineering, Kumoh National Institute of Technology, Gumi, 39177, Korea

*Corresponding Author: Sungyoung Kim. Email: sykim@kumoh.ac.kr

Received: 18 September 2023 Accepted: 28 November 2023

ABSTRACT

In the context of high compression rates applied to Joint Photographic Experts Group (JPEG) images through lossy compression techniques, image-blocking artifacts may manifest. This necessitates the restoration of the image to its original quality. The challenge lies in regenerating significantly compressed images into a state in which these become identifiable. Therefore, this study focuses on the restoration of JPEG images subjected to substantial degradation caused by maximum lossy compression using Generative Adversarial Networks (GAN). The generator in this network is based on the U-Net architecture. It features a new hourglass structure that preserves the characteristics of the deep layers. In addition, the network incorporates two loss functions to generate natural and high-quality images: Low Frequency (LF) loss and High Frequency (HF) loss. HF loss uses a pretrained VGG-16 network and is configured using a specific layer that best represents features. This can enhance the performance in the high-frequency region. In contrast, LF loss is used to handle the low-frequency region. The two loss functions facilitate the generation of images by the generator, which can mislead the discriminator while accurately generating high- and low-frequency regions. Consequently, by removing the blocking effects from maximum lossy compressed images, images in which identities could be recognized are generated. This study represents a significant improvement over previous research in terms of the image resolution performance.

KEYWORDS

JPEG; lossy compression; restoration; image generation; GAN

1 Introduction

Color image compression can be conceptually analogous to the compression of multiple monochromatic channels, where each channel is either encoded sequentially in its entirety or encoded through a method of alternatingly multiplexing blocks of samples (such as 8×8 blocks) from each respective channel. The JPEG (Joint Photographic Experts Group) standard is a technology for compressing continuous-tone images and aims to accommodate a variety of source image formats [1]. JPEG utilizes lossy compression technology to remove particular image data and reduce the file size intentionally. Only parts irrelevant to human vision are removed during this process. The JPEG image compression process follows a series of steps. It starts with color-space conversion to

eliminate irrelevant image content. At this stage, RGB images are converted to the YCbCr color space because the human eye is more sensitive to brightness components than to color components. The subsequent step involves performing 4:2:0 subsampling to further reduce the file size. During subsampling, the Cb and Cr components are sampled while Y is preserved. In the context of the JPEG compression procedure, color transformation and subsampling are often implemented, yet they are not officially part of the JPEG specification. The JPEG format itself does not provide a complete image representation. The JPEG File Interchange Format (JFIF) complements this by specifying parameters for sample alignment, resolution, aspect ratio, and color space in the container format of JPEG-encoded data. After subsampling, each component image undergoes blockwise division into fixed-sized blocks. Moreover, a discrete cosine transform (DCT) is applied to each block to calculate its spatial frequency. The spatial frequency represents the rate of variation in the intensity or color of an image with respect to distance or space. Thereby, it signifies how frequently the image intensity or color varies within a given spatial distance. A high spatial frequency indicates the presence of intricate, detailed patterns originating from numerous variations in intensity or color within a specific distance. Meanwhile, a low spatial frequency signifies a smooth or coarse pattern originating from fewer fluctuations in intensity or color within the same distance. After block-wise Discrete Cosine Transform (DCT), the coefficients are quantized using a quantization matrix, an input parameter required by the encoding application. Quantization aims to enhance compression efficiency by encoding DCT coefficients with the minimal precision required to maintain the desired image quality threshold. This is followed by zigzag scanning to transform these into one-dimensional data. Finally, the DC coefficients are encoded based on differential pulse code modulation (DPCM) and Huffman encoding, whereas the AC coefficients are encoded using run length and Huffman encoding. This series of processes culminates in the final compressed image.

Blockwise quantization in the DCT can result in blocking artifacts when the compression rate is high. In our prior research [2], which is relevant to this study, we conducted an analysis by converting PNG images into JPEG formats with various compression rates. The analysis revealed that as the compression ratio exceeded 95% (20:1), the blocking artifacts became more observable. The objective of this study was to restore images that have been compressed with a maximum compression ratio of 98% (50:1) using the JPEG algorithm by employing multiple loss functions and developing new deep learning network structures. In another preliminary study [3], we evaluated the feasibility of restoring a maximum compressed image using an image-to-image method. Fig. 1 illustrates that employing the highest compression rate of 98% results in the emergence of visible blocking artifacts. This hinders the discernment of the image content. Furthermore, substantial degradation of color information occurs, thereby rendering the faces unidentifiable. Restoration of the original image content using maximum compression provides a substantial advantage in terms of image security.

Our main contribution is as follows:

- We explore methods for restoring images that have undergone maximum lossy compression in JPEG format. While previous studies have addressed restoration from other formats or partially compressed images, the challenge of restoring from maximum lossy compression of JPEG has not yet been tackled. Images in this state possess significantly reduced information, making restoration especially challenging. In this paper, we propose a novel approach to overcome this issue.
- The U-Net network is widely recognized for its superior performance in the realms of image segmentation and restoration. Building upon the robust foundation of U-Net, we introduced a novel "hourglass block" to enhance the generation capabilities of latent vectors. The introduced hourglass block enhances the already notable feature extraction capabilities, leading to the

formation of more precise and stable latent vectors. The integration of this novel component with U-Net established architecture has played a pivotal role in significantly elevating the precision and stability of image restoration tasks.

- In the conventional GAN-based training approach, the capability of the generator to restore images is severely limited because of the low amount of information available in the images. This results in a high classification performance of the discriminator. The introduction of an early stopping discriminator to address this limitation yielded a performance improvement.

Figure 1: Examples of JPEG lossy compressed images with maximum compression ratio of 98%

2 Related Work

2.1 Lossy Compression

Mentzer et al. [4] introduced a method for lossless compression using the Better Portable Graphics (BPG) format, which is an algorithm for lossless image compression. They compressed original images with BPG and then decomposed the residuals into reconstructions to model the distribution of residuals using a CNN-based approach. Their results, which utilized residual coders and entropy coding, demonstrated high compression efficiency. Qin et al. [5] proposed a lossy compression method for encrypted images. Their method involved encryption using modulo-256 arithmetic, compressing the image while avoiding pixel distortion through image inpainting techniques. Yan et al. [6] developed a framework based on Generative Adversarial Networks (GANs) to achieve the lowest Mean Squared Error (MSE) distortion at given bit rates. Their study provided theoretical evidence that distortion loss has a negligible impact on the training process.

2.2 Image Generation

DCGAN [7] marks the evolution of generation models, building on the foundation of the original GAN [8]. This model emerged from significant efforts including the adoption of a CNN-based structure in place of fully connected layers and a revision of the activation function. It notably demonstrated the ability to generate variations through 'walking in the latent space' by manipulating the latent vector. Pix2Pix [9], another GAN-based network, derives from cGAN [10]. Unlike models relying on latent vectors, Pix2Pix processes an input image of one style and outputs it in another. This supervised learning approach requires both input data and corresponding target images. PGGAN [11] enhances image resolution by starting with low-resolution training images and incrementally increasing image size by adding layers. This methodology enables stable generation of high-resolution images and reduces training time by focusing on details in a stepwise manner. Building on PGGAN [11], StyleGAN [12] incorporates style transfer into the network structure, allowing for specific scale control not achievable in PGGAN. This is facilitated by disentangling the latent vector and quantifying it. Additionally, AdaIN [13] is utilized to ensure that each scale is influenced by a unique style. However, this approach exhibited flaws like water droplet-like artifacts, especially at higher resolutions. StyleGAN2 [14] addresses this by normalizing the mean and variance of each AdaIN [13] layer, thereby disrupting feature interdependencies. StyleGAN-ADA [15] is a progression from StyleGAN [12] and StyleGAN2 [14], offering a solution to discriminator overfitting in cases of limited dataset size. It involves training the generator and discriminator with augmented images, adjusting the augmentation range based on probability. The field of image generation continues to advance rapidly, with numerous applications ranging from generating human postures from key points [16], predicting future frames from multiple frames [17], to removing clouds from solar images [18].

2.3 Super Resolution

Super-resolution research primarily focuses on converting low-resolution images into high-resolution images to enhance image quality. While the content of this paper technically does not fall under super-resolution, it shares the common objective of improving image quality with super-resolution research. However, a key difference is that super-resolution research does not use compressed images as inputs, unlike this paper.

SRCNN [19] is a CNN employing an end-to-end mapping method to restore low-resolution images. It stands out for its simplicity and speed, relying solely on convolutional layers. ESPCN [20] introduces a sub-pixel layer to counter the computational load increase associated with upscaling at the network's input. This layer allows the network to process images as if they are smaller by rearranging pixels in the final feature map, enhancing performance. VDSR [21] uses a deep network with 20 layers to leverage the texture information from the entire image. It also shows speed improvements through adjustable gradient clipping. SRGAN [22] enhances image quality using a GAN-based structure, overcoming previous limitations in restoring high-frequency area information. Its approach to generating more visually appealing images through high-level feature maps from the VGG network aligns with the methods discussed in this study. ESRGAN [23], an extension of SRGAN [22], further improves the realism and naturalness of images through various techniques, including the introduction of RRDB and utilizing pre-activation features.

2.4 Quality Enhancement through Artifact Removal

HiFaceGAN [24] introduces an algorithm that employs multiple CSR units for hierarchical analysis and restoration of image features. Each unit targets different feature levels, following a

sequential restoration process from low-resolution inputs. This model is trained on a range of image degradations, particularly focusing on eliminating artificial compression artifacts and improving the quality of JPEG images. Although similar in approach to prior research, a notable distinction of this method is its lack of focus on extreme compression scenarios. ESDNet-L [25] offers a solution for removing moiré patterns, addressing the aliasing phenomenon that impairs image quality. This study successfully demonstrates the removal of moiré patterns from high-resolution 4K images while maintaining rapid processing speeds. While its primary objective is aliasing removal, the method also shows relevance in addressing blocking artifacts common in heavily compressed JPEG images, indicating some parallels with the proposed approach. However, a unique aspect of this study is its emphasis on ultra-high resolution, differentiating it from other methods.

3 Restoration of Maximum Lossy Compression Image

The objective of this study was to restore maximum lossy compression images. This is a challenging task because of the difficulty in recovering data with limited information, as depicted in Fig. 1. However, the proposed GAN-based structure demonstrated a high performance in restoring such images to their original state. The overall learning structure of this study is illustrated in Fig. 2. The original image is denoted by X and defined as $C(X)$ by generating the maximum lossy compressed image using the JPEG compressor. The supervised learning method is used in the form of a pair between X and $C(X)$. Here, the training data configuration can be expressed as $S_{data} = \{(X_i, C(X_i)), X_i \in X, C(X_i) \in C, i = 1, 2, \ldots, N\}$, and the size of the image is (128, 128, 3). The input to the generator is denoted by $C(X)$, and its output is denoted by $G(C(X))$. The objective of the generator is to restore the original image X (which has undergone maximum lossy compression) such that it closely resembles the original content. In the GAN architecture, the generators produce images closer to x to mislead the discriminator, whereas the discriminator is trained to accurately differentiate between $G(C(X))$ and X. The generator and discriminator have a minimax structure that conflicts with the generation of more natural images. This results in reconstructed images that are closer to the original image.

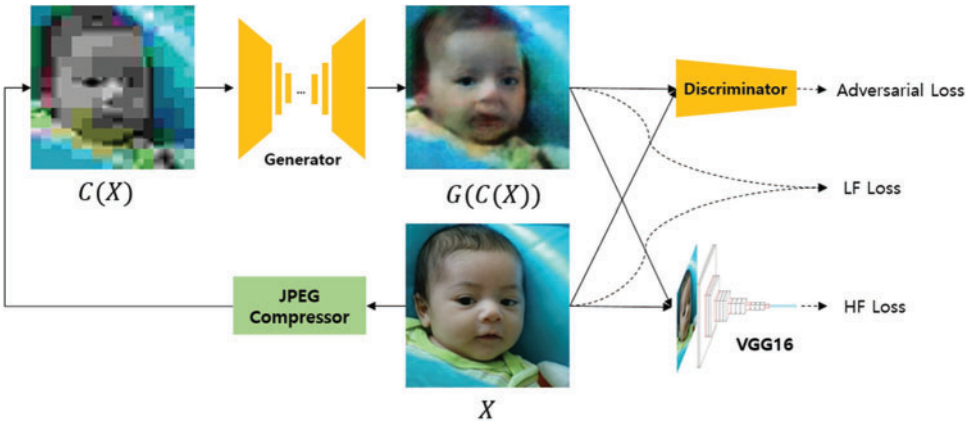

Figure 2: Overall training architecture

3.1 Generator

The U-Net-based network structure of the generator in this study is illustrated in Fig. 3. It was designed to restore the maximum lossy compressed images. Unlike the existing U-Net, it has a depth of seven with the input and output designed to be of an equal size. The input and output of the network

were color images of (512, 512, 3), whereas the original image size was (128, 128, 3). Encoding extracts features in six layers, and decoding restores these to their original size. In addition, an hourglass network structure is placed in the middle of the network to extract small features that were lost in the dataset. This structure is repeated four times at the depths of the 6th and 7th layers to extract effective features. All the layers have a 4×4 filter and stride of two. The depths of the 6th and 7th layers are set to 1024 to extract many features. The activation function for all the layers except the final decoding layer is leaky-ReLU. The final layer uses tanh. Finally, the decoding result is brought to its original size to generate a natural image. However, linear interpolation is used to adjust the image size from (128, 128, 3) to (512, 512, 3). This increases the number of pixels by 16 times, which may result in a longer training time.

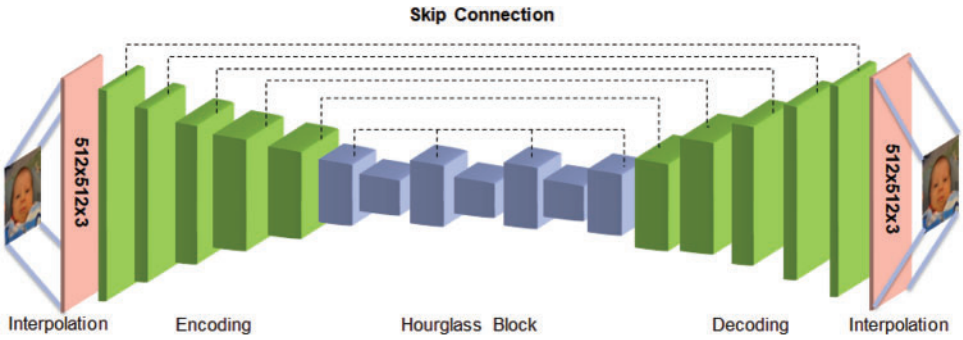

Figure 3: Generator architecture

3.2 Discriminator

In this study, the architecture of the discriminator network is based on the PatchGAN [26] structure, as depicted in Fig. 4. The input of the discriminator was formed by concatenating x and $G(C(X))$. This resulted in a feature map of size (62, 62, 1). It corresponds to the receptive field proposed by the PatchGAN [26] and is used to train the discriminator to classify the generated $G(C(X))$ as false. Layers 1–3 use leaky-ReLU as the activation function, and the stride is fixed at two. In the fourth layer, an activation function using leaky-ReLU is applied, and the stride is fixed at one with padding. In the fifth layer, the activation function uses a sigmoid function to produce an output between zero and one, and the structure is identical to the fourth layer.

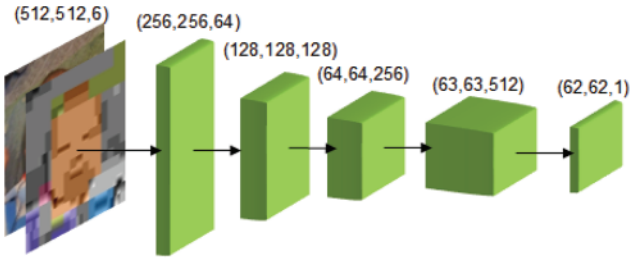

Figure 4: Discriminator architecture

3.3 Loss Function

The proposed loss function comprises three main functions. However, when the log function commonly used in the existing GAN is applied, the problem of dimensionality occurs in the proposed method. This is explained in Section 4. To achieve stable learning without this problem, the log function was removed, and the loss was defined only by the output of the existing discriminator. The Discriminator is trained to classify $D(G(C(X)))$ as close to 0 and $D(X)$as close to 1, resulting in an Adversarial loss output value between −1 and 1, which is trained to minimize according to Eq. (1). The aim of the generator is to make $G(C(X))$ appear real to the discriminator. Therefore, the loss is set to minimize Eq. (2). Consequently, Eqs. (1) and (2) have converse structures and can be used in the GAN training process.

$$L_{Adv}(D) = \mathbb{E}_{X_i \sim X, C(X_i) \sim C}[D(G(C(X_i))] - \mathbb{E}_{X_i \sim X}[D(X_i)] \tag{1}$$

$$L_{Adv}(G) = \mathbb{E}_{X_i \sim X, C(X_i) \sim C}[1 - D(G(C(X_i))] \tag{2}$$

The images can be primarily divided into two components based on their elements: high and low frequencies. The high-frequency region includes the edges and details of the object, whereas the low-frequency region includes the color and texture of the object. In the case of facial images, the high-frequency region may correspond to an individual's identity, whereas the low-frequency region may be referred to as the excluded part. To preserve the low-frequency region, the proposed loss function employs the L1 loss (as shown in Eq. (3)) by minimizing the difference between the pixel values of the generated and original images. To preserve the high-frequency areas, a pretrained VGG-16 model is utilized. The feature maps that can effectively preserve the high-frequency regions are selected using the L2 loss (as defined in Eq. (4)), which is referred to as the HF loss.

$$L_{LF} = \mathbb{E}_{X_i \sim X, C(X_i) \sim C} |X_i - G(C(X_i))| \tag{3}$$

$$L_{HF} = \mathbb{E}_{X_i \sim X, C(X_i) \sim C}[l_i(X_i) - l_i(G(C(X_i)))]^2 \tag{4}$$

The total loss function (expressed in Eq. (5)) is the sum of the three loss functions with the aim of minimizing these. The hyperparameters are set according to the weight of each loss function, as described in Section 4.

$$L_{Total} = \lambda_{Adv} L_{Adv}(G, D) + \lambda_{LF} L_{LF} + \lambda_{HF} L_{HF} \lambda_{HF} L_{HF} \tag{5}$$

4 Experimental Results

4.1 Datasets

The datasets used in the training and testing phases were obtained from the FFHQ datasets proposed by StyleGAN [12]. FFHQ is a collection of western facial datasets containing 4,000 images partitioned in the ratio 8:2 for training and testing purposes. The images are in the PNG format, which provides lossless compression, thereby preserving the image quality during compression. Additionally, it supports the alpha channel, allowing for the preservation of images. Consequently, it retains rich color information, ensuring better depth and fidelity in color details. To generate the maximum number of lossy compressed images (JPEG compressor), the PNG images were converted to JPEG and compressed. This resulted in a compressed data rate of approximately 96.7%. The maximum lossy compressed images and original FFHQ images are color images with sizes of (128, 128, 3) and are paired for evaluation.

4.2 Training Details

All experiments were conducted on the Ubuntu 18.04 LTS operating system using GeForce RTX 3090. The overall training process utilized the stochastic gradient descent method with batch size 1. The training data consisted of 3,200 images, and with 30 epochs, it trained for a total of 96,000 iterations. The original image size was up-scaled to (512,512,3) using linear interpolation as an input to the model. After obtaining output, it was then down-scaled back to the original size, and evaluated using a loss function. This strategy allows for including more information within the feature maps during training. For every layer in the network, operations are followed by the application of batch normalization and an activation function. For batch normalization, an epsilon value of 1e-5 and a momentum of 0.9 are consistently used throughout the experiments. The activation function employed is the leaky ReLU. The structure of the generator is as shown in Table 1. The encoder and decoder have mirror-image sizes, and the hourglass block is repeated a total of four times. Both generator and discriminator employed the Adam optimizer with an initial learning rate of 0.00001. Additionally, a dropout rate of 70% was applied to the hourglass blocks. Because lossy compressed images contain limited information, generating high-quality images is challenging, particularly for generators. Consequently, the discriminator cannot effectively distinguish between real and counterfeit images. This hinders the production of natural and high-quality images. To overcome this challenge, the discriminator was not trained after 10 epochs, and a higher-quality image was generated. With λ_{Adv} set to one and low-frequency images comprising most of the data, λ_{LH} was set at a high value of 20. In contrast, the high-frequency images underwent training with λ_{FH} set at 0.1. To detect the high-frequency region, only Conv4_1 of VGG-16 was used. This preserved the identity of the number of individuals. Finally, the maximum lossy compressed image was restored, and the generated image was saved in a PNG file format similar to the original image.

Table 1: Detailed generator architecture of proposed method

	Layer	Input size	Output size	Activation functions
Encoder	1	[N, 512, 512, 3]	[N, 256, 256, 64]	Leaky-ReLU
	2	[N, 256, 256, 64]	[N, 128, 128, 128]	Leaky-ReLU
	3	[N, 128, 128, 128]	[N, 64, 64, 256]	Leaky-ReLU
	4	[N, 64, 64, 256]	[N, 32, 32, 512]	Leaky-ReLU
	5	[N, 32, 32, 512]	[N, 16, 16, 512]	Leaky-ReLU
	6	[N, 16, 16, 512]	[N, 8, 8, 512]	Leaky-ReLU
Hourglass Block (Repeat 4)		[N, 8, 8, 512]	[N, 4, 4, 512]	Leaky-ReLU
		[N, 4, 4, 512]	[N, 8, 8, 512]	Leaky-ReLU
Decoder	1	[N, 8, 8, 512]	[N, 16, 16, 512]	Leaky-ReLU
	2	[N, 16, 16, 512]	[N, 32, 32, 512]	Leaky-ReLU
	3	[N, 32, 32, 512]	[N, 64, 64, 256]	Leaky-ReLU
	4	[N, 64, 64, 256]	[N, 128, 128, 128]	Leaky-ReLU
	5	[N, 128, 128, 128]	[N, 256, 256, 64]	Leaky-ReLU
	6	[N, 256, 256, 64]	[N, 512, 512, 3]	Tanh

4.3 Results Analysis

The results of the proposed method are shown in Fig. 5. The compressed images are shown in Figs. 5a and 5d , and the original images are shown in Figs. 5c and 5f. Figs. 5b and 5e show the images generated by the generator. Overall, the generator produced an image that was close to the original image, using a compressed image with highly marginal information. Specifically, there was a significant loss of skin and hair color. However, it was demonstrated that even these parts could be regenerated. Meanwhile, in the case of maximum lossy compression, the residual image data were restricted significantly. This resulted in the successful restoration of the overall structure while inadequately capturing intricate image details. This limitation originates from the incapability of the compressed image to preserve fine elements such as wrinkles, whiskers, and hair, thereby rendering these imperceptible to human observers. For example, wrinkles cannot be recreated effectively. This causes the generated image to portray a youthful appearance and generally results in closed-mouth images displaying teeth. This phenomenon occurs because of the inadequacy of information within the compressed data (which predominantly consists of smiling facial expressions), which induces the generator to produce open-mouth renditions. Consequently, the process of restoration using maximum lossy compressed images exhibits inherent constraints. The approach proposed in this study addresses these constraints and demonstrates its capability to generate natural and high-quality images.

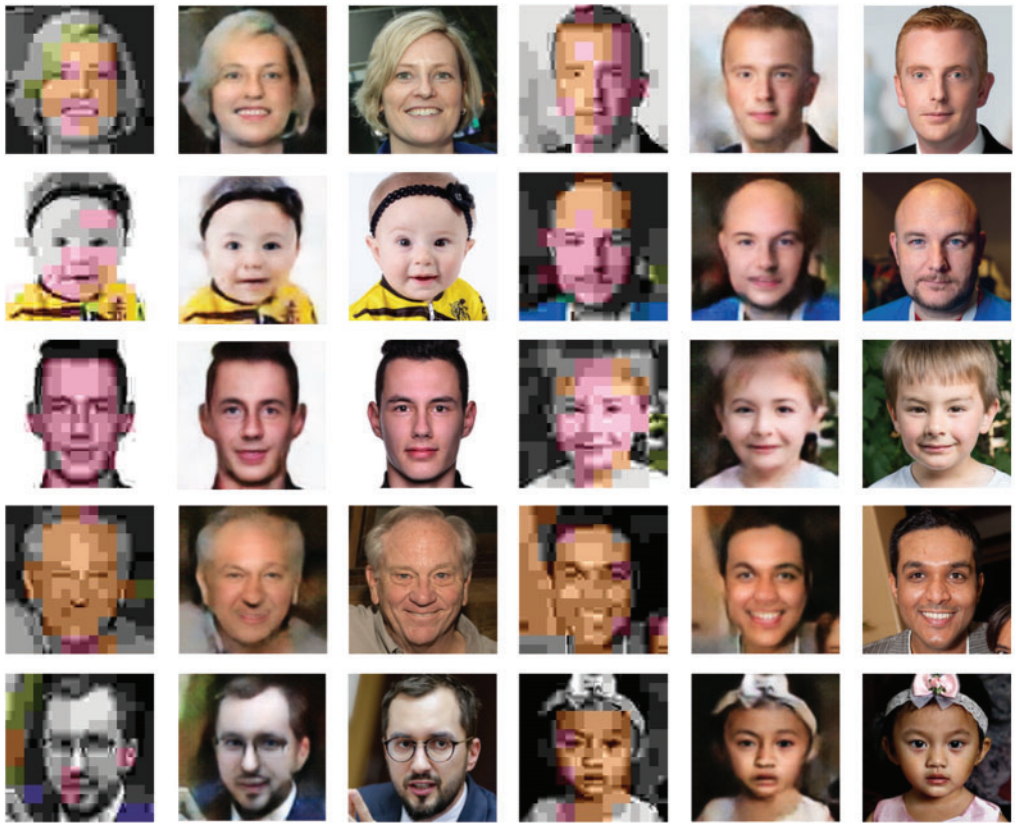

Figure 5: (Continued)

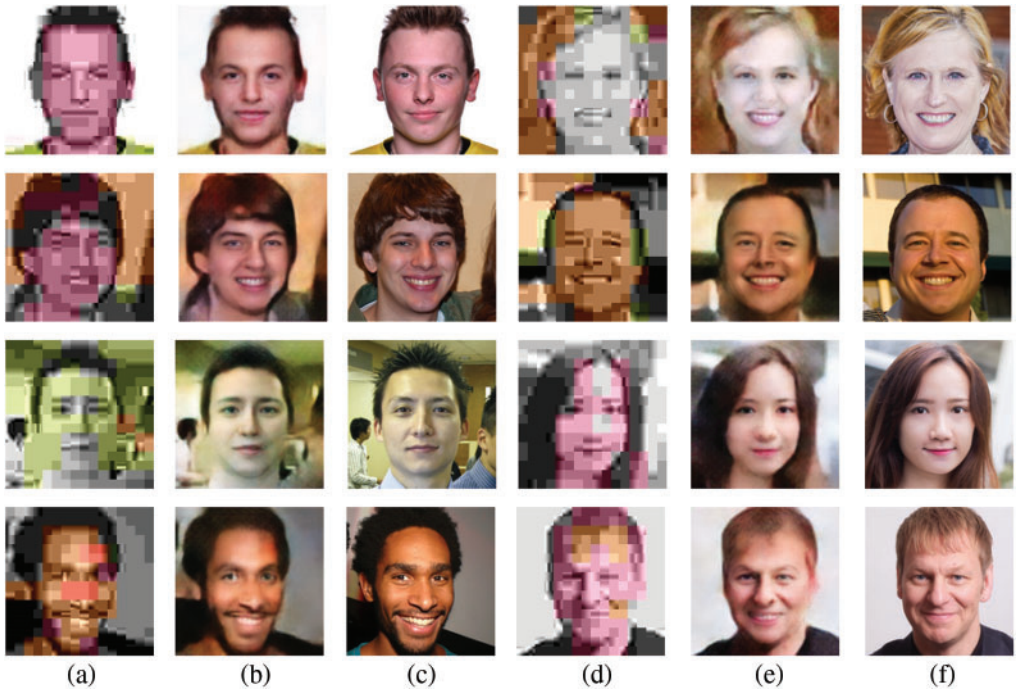

Figure 5: Comparison of images ((a,d): compressed images, (b,e): restored images, (c,f) GT)

4.4 Performance Evaluation

For image quality evaluation, the performance was assessed based on commonly used metrics such as PSNR and SSIM [26]. In addition, further evaluation was performed using VIF [27] and LPIPS [28].

The PSNR is a quality index that can measure the loss of information of an image owing to the loss of quality. The better the result, the smaller is the difference from a pixel perspective. However, because it considers only the pixel perspective, it can evaluate the low-frequency area but does not consider the human visual aspect. Thus, a high score could be obtained even if a natural image is not generated. The equation for the PSNR is presented in Eq. (6). A higher score corresponds to a better performance.

$$\mathrm{PSNR}(X, G(C(X))) = 10log\left(\frac{Max^2}{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(X(i,j) - G(C(X))(i,j))^2}\right) \tag{6}$$

The SSIM is a widely used image quality measurement index in engineering that considers human visual content. It evaluates the similarity between the original and generated images based on factors such as structure, brightness, and contrast. The closer the SSIM value is to one, the more similar are the two images. SSIM is a more reliable method than PSNR because it overcomes the shortcomings of PSNR, which considers only pixel differences. The SSIM is calculated using the formula presented in Eq. (7):

$$\mathrm{SSIM}(X, G(C(X))) = \frac{(2\mu_{xX}\mu_{G(C(X))} + C_1)(2\sigma_{X,G(C(X))} + C_2)}{(2\mu_X^2 + \mu_{G(C(X))}^2 + C_1)(\sigma_X^2 + \sigma_{G(C(X))}^2 + C_2)} \tag{7}$$

VIF [27] is a metric that measures the fidelity of an entire reference image to evaluate its quality. It performs this measurement by comparing the information content of the original and generated images. This indicator uses the results of the human visual system and a Gaussian distribution to assess the image quality. The high information content in the generated image even after being input as a lost compressed image is indicative of a good performance. The performance evaluation is expressed using Eq. (8). Here, a VIF value closer to one indicates a higher performance.

$$\mathrm{VIF}(X, G(C(X))) = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} log\left(1+\frac{g_i^2 s_i^2 \lambda_j}{\sigma_X^2+\sigma_{G(C(X))}^2}\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W} log\left(1+\frac{s_i^2\lambda_j}{\sigma_X^2}\right)} \tag{8}$$

LPIPS (Learned Perceptual Image Patch Similarity) [28] is a metric designed to measure the perceptual similarity between two images, leveraging feature maps extracted from ImageNet classification models. By assessing the similarity between activation maps across multiple layers, it offers an evaluation closely aligned with human perception. A lower LPIPS score indicates that the two images are more similar, meaning that models demonstrating high performance will exhibit lower LPIPS values.

A variety of performance comparisons are presented in Table 2. The performance of the proposed method was better than those of previous studies. The study also evaluated the performance of waifu2x [29] and preliminary study [3], which improved the resolution of CNN-based 2D images. The preliminary study [3] is carried out to achieve the same goal as the method being proposed, based on Pix2Pix [8]. HifaceGAN [24] is evaluated using a model trained for the purpose of JPEG Compression Artifact Removal. In the study conducted by ESDNet-L [25], the efficacy of a deep learning model in removing blocking artifacts and enhancing the clarity of images subjected to severe lossy compression is examined.

Table 2: Performance comparison including maximum lossy compressed images

	PSNR ↑	SSIM ↑	VIF ↑	LPIPS ↓
Lossy compression	20.381(0.704)	0.582(0.002)	0.201(0.001)	0.670(0.003)
Preliminary study [3]	20.537(0.857)	0.586(0.002)	0.211(0.001)	0.503(0.002)
waifu2x [29]	20.581(0.466)	0.599(0.003)	0.211(0.001)	0.639(0.002)
ESRGAN [23]	19.763(0.714)	0.539(0.003)	0.180(0.001)	0.666(0.002)
HiFaceGAN [24]	20.591(0.703)	0.597(0.002)	0.207(0.001)	0.665(0.003)
ESDNet-L [25]	17.263(4.308)	0.573(0.003)	0.199(0.001)	0.617(0.002)
Ours	**21.680(1.436)**	**0.675(0.003)**	**0.261(0.001)**	**0.369(0.003)**

This study compared the generation results of previous studies with the proposed method in Fig. 6. $C(x)$ refers to the lossy compressed image that exhibits blocking artifacts in Fig. 6a. However, after the application of waifu2x [29] (Figs. 6b–6d), there was no significant difference from the original image. Although super-resolution methods are generally necessary for evaluating the performance of a study, the purpose of this research was to restore a lossy compressed image rather than to improve its quality. Therefore, existing image-quality improvement methods do not exert a significant effect. Fig. 6e displays a visually distinct result as results of the research to remove JPEG blocking, different

from the previously mentioned results. However, it presents a blurred outcome to the extent that identity verification becomes impossible, indicating a suboptimal result. The method proposed in Fig. 6f produces more natural and clearer results than those of previous studies. Results of our method display a quality very similar to ground truth (Fig. 6g), clear enough to discern the identity.

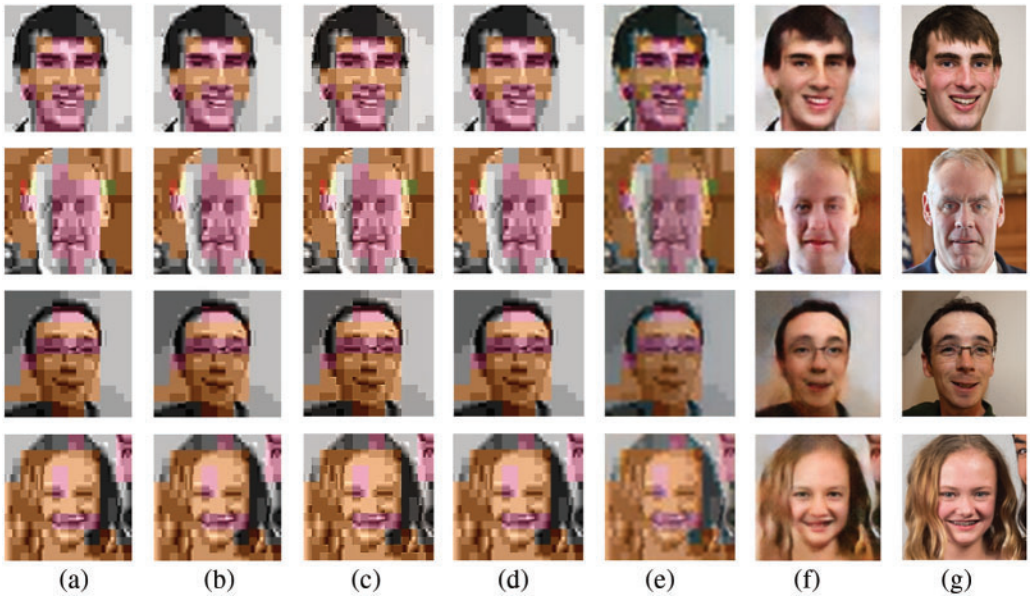

Figure 6: Comparison with related works ((a): lossy-compressed image, (b): waifu2x [28], (c): ESRGAN [23], (d): HiFaceGAN [24], (e): ESDNet-L [25], (f): our method, (g): GT)

4.5 Ablation Study

The concept of a GAN is a critical aspect of this research because it enhances the performance of the generator and discriminator modules compared with the use of only convolutional neural network. The ultimate objective of a GAN is for the generator to produce highly realistic data and for the discriminator to distinguish between real and generated images. However, the generator operates based on the maximum lossy compressed image. This enables the generation of a realistic image that can mislead the discriminator. This is owing to the minimal amount of information in the maximum lossy compressed image. This facilitates the discriminator in distinguishing between the real and generated images. Although the GAN structure may work effectively in the early stages before the discriminator is trained, the generator's focus on beating the discriminator rather than training for good performance prevents it from generating high-quality images. This study proposed a new training method in which the discriminator is trained for only 10 epochs. Then, the generator is trained to optimize a loss function other than beating the discriminator. Table 3 presents the results of the comparative analysis to determine whether to alter λ_{LH}, which is significantly involved in generating low frequencies and attracting the discriminator to 10 epochs. For metrics such as PSNR, SSIM, and VIF, optimal results were achieved when λ_{LH} was set to 20 and the process was stopped. However, for LPIPS, the optimal setting was at 5. The difference between the stopped and not stopped scenarios in the LPIPS metric was less than 0.1, making it challenging to distinguish the performance. In contrast, the PSNR, SSIM, and VIF metrics exhibited a significant performance difference. Excluding LPIPS, all other metrics demonstrated superior performance when stopped.

Table 3: The results of the experiment according to not stop or stop of discriminator training and hyperparameter

	$\lambda_{LH} = 5$		$\lambda_{LH} = 10$		$\lambda_{LH} = 20$	
	Not Stop	Stop	Not Stop	Stop	Not Stop	Stop
PSNR ↑	21.558	21.617	21.573	21.675	21.596	**21.680**
SSIM ↑	0.666	0.670	0.670	0.673	0.671	**0.675**
VIF ↑	0.255	0.257	0.257	0.259	0.257	**0.261**
LPIPS↓	0.356	**0.355**	0.362	0.362	0.368	0.369

The contribution of this study was assessed by evaluating the performance of the model with and without the hourglass block structure and VGG-16 in Table 4. The results indicated that without the two, the performance was low in all the aspects compared with the highest performance. The method proposed in this paper is a marginally improved version of preliminary study [3]. The addition of the hourglass block structure improved the performance. This demonstrated that the block can aid in maximizing the representation of the conserved portion of the high-dimensional feature map. The addition of high-frequency region preservation using VGG-16 yielded the highest performance mentioned in this paper. The proposed method preserves natural images and maintains the identity of the individual. This results in a relatively high-quality image.

Table 4: Comparison of performance according to the use of hourglass block and VGG-16

Hourglass block	X	O	O
VGG-16	X	X	O
PSNR ↑	20.623	21.445	**21.680**
SSIM ↑	0.604	0.665	**0.675**
VIF ↑	0.222	0.255	**0.261**
LPIPS ↓	0.398	0.398	**0.369**

Table 5 illustrates the performance comparison based on the number of hourglass blocks proposed in this paper. Using just one is equivalent to the existing U-Net structure, and thus, is excluded from the evaluation. The table presents the performance for 2 to 4 repetitions. Overall, repeating 4 times yielded the highest results when measured using PSNR, SSIM, VIF, and LPIPS. As the number of repetitions increases, the performance in terms of PSNR, SSIM, and VIF gradually improves. However, the performance evaluation results for LPIPS were consistent between 2 and 4 repetitions. Based on the performance metrics used for evaluation, repeating the process 4 times is considered the most optimal.

4.6 Limitations

Because this study focused on restoration based on maximum lossy compression images, the restoration process is limited in its capability to produce exact replicas. However, the objective was to generate datasets that closely resemble the original dataset through deep learning network-based learning. Although certain restorations are feasible at the general shape level, the details cannot be restored. Although similar colors can be generated, wrinkles and hair features cannot be reproduced.

Moreover, the restoration performance is inferior for the dataset on the face side, as shown in Fig. 7. This is because the majority of the datasets used for restoration have a forward-facing orientation, and the restoration technique restores the missing content based on its characteristics. As shown in the restored image in Fig. 7, the opposite eye is generated notwithstanding the absence of any relationship with the frontal data. Additionally, the face is blurred. To improve the restoration performance of side images, it is necessary to include side-facing images in the training data.

Table 5: Comparison of performance according to the number of hourglass block

Hourglass block	2	3	4
PSNR ↑	21.559	21.639	**21.680**
SSIM ↑	0.668	0.672	**0.675**
VIF ↑	0.255	0.257	**0.261**
LPIPS ↓	**0.369**	0.371	**0.369**

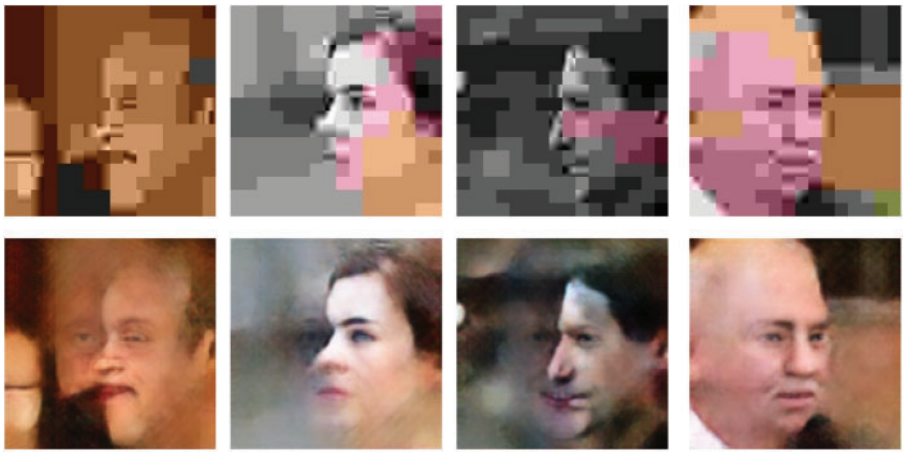

Figure 7: The results on the side of face images (top: compressed images, bottom: results)

5 Conclusions

This paper presents a GAN-based network approach for restoring JPEG images using data with maximum lossy compression. Notwithstanding the difficulty of restoring images with minimal information, even to the human eye, the proposed method demonstrated satisfactory performance. Although certain limitations exist in achieving precise restoration, this method was demonstrated to restore images to a certain degree. Future work is likely to further improve the restoration performance. This could, in turn, potentially expand the application of lost compressed image restoration to the field of image security.

Acknowledgement: I would like to express gratitude to my advisor for the invaluable guidance and support in preparing this paper.

Funding Statement: This work was supported by the Technology Development Program (S3344882) funded by the Ministry of SMEs and Startups (MSS, Korea).

Author Contributions: Study conception and design: J. Si, S. Kim; data collection: J. Si; analysis and interpretation of results: J. Si, S. Kim; draft manuscript preparation: J. Si, S. Kim.

Availability of Data and Materials: In this study, we used a public dataset, which can be downloaded from the website if needed (https://github.com/NVlabs/ffhq-dataset).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. 18–34, 1992.

[2] J. Si and S. Kim, "Quality analysis of the conversion according to the compression rate from the PNG image to the JPEG image," in *Proc. of Korean Institute of Information Technology (KIIT)*, Jeju, Korea, pp. 109–111, 2021.

[3] J. Si and S. Kim, "Restoration of JPEG lossy compressed image based on Pix2Pix: A Preliminary Study," in *Proc. of Korean Institute of Information Technology (KIIT)*, Jeju, Korea, pp. 53, 2022.

[4] F. Mentzer, L. V. Gool and M. Tschannen, "Learning better lossless compression using lossy compression," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 6638–6647, 2020.

[5] C. Qin, Q. Zhou, F. Cao, J. Dong and X. Zhang, "Flexible lossy compression for selective encrypted image with image inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3341–3355, 2019.

[6] Z. Yan, F. Wen, R. Ying, C. Ma and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework," in *Proc. of Int. Conf. on Machine Learning*, pp. 11682–11690, 2021.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of Int. Conf. on Learning Representations (ICLR)*, San Diego, USA, pp. 1–14, 2015.

[8] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley *et al.,* "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1–9, 2014.

[9] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 1125–1134, 2017.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784v1, 2014.

[11] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, abd variation," in *Proc. of Conf. on Learning Representations (ICLR)*, Vancouver, Canada, pp. 1–26, 2018.

[12] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long beach, USA, pp. 4401–4410, 2019.

[13] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 1501–1510, 2017.

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen *et al.,* "Analyzing and improving the image quality of StyleGAN," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020.

[15] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen *et al.,* "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12104–12114, 2020.

[16] A. Siarohin, E. Sangineto, S. Lathuilière and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 3408–3416, 2018.

[17] J. Si and S. Kim, "Traffic accident detection in first-person videos based on depth and background motion Estimation," *Journal of Korean Institute of Information Technology (JKIIT)*, vol. 19, no. 3, pp. 25–34, 2021.
[18] X. Wu, W. Song, X. Zhang, G. Lin, H. Wang *et al.,* "Algorithm development of cloud removal from solar images based on Pix2Pix Network," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3497–3512, 2022.
[19] C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.
[20] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken *et al.,* "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 1874–1883, 2016.
[21] J. Kim, J. Lee and K. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 1646–1654, 2016.
[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham *et al.,* "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 4681–4690, 2017.
[23] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Proc. of European Conf. on Computer Vision (ECCV)*, Amsterdam, Netherlands, pp. 702–716, 2016.
[24] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu *et al.,* "HiFaceGAN: Face renovation via collaborative suppression and replenishment," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, New York, USA, pp. 1551–1560, 2020.
[25] X. Yu, P. Dai, W. Li, L. Ma, J. Shen *et al.,* "Towards efficient and scale-robust ultra-high-definition image demoiréing," in *European Conf. on Computer Vision (ECCV)*, Tel Aviv, Israel, pp. 646–662, 2022.
[26] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. of Int. Conf. on Pattern Recognition (ICPR)*, Istanbul, Turkey, pp. 2366–2369, 2010.
[27] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp. 586–595, 2018.
[29] waifu2x. 2023. [Online]. Available: http://waifu2x.udp.jp/index.ko.html (accessed on 25/09/2023).